The Journal of Computer Science and Its Applications. An International Journal of the Nigeria Computer Society (NCS)
Vol.32, No.1, June, 2026

Algorithms for Crime Type Comparative Analysis of Supervised Machine Learning Detection and Prediction

***Abass Olalere A.**
Dept. of Computer Sciences,
Olabisi Onabanjo University,
Ago-Iwoye, Nigeria.
abass.olalere@oouagoiwoye.edu.ng

Alamu Babatunde M.
Dept. of Computer Sciences,
Tai Solarin Federal University of Education,
Ijagun, Nigeria.
alamumichael15@gmail.com

Ogunsanwo Gbenga O.
Dept. of Computer Sciences,
Tai Solarin Federal University of Education,
Ijagun, Nigeria.
ogunsanwogo@tasued.edu.ng

Alaba Olumuyiwa B.,
Dept. of Computer Sciences,
Tai Solarin Federal University of Education,
Ijagun, Nigeria.
alabaob@tasued.edu.ng

Omude Paul O.
Dept. of Computer Sciences,
Tai Solarin Federal University of Education,
Ijagun, Nigeria.
omudepaul@gmail.com

Olukoya Gabriel O.
Dept. of Computer Sciences,
Tai Solarin Federal University of Education,
Ijagun, Nigeria.
olukoyaisrael@gmail.com

*Corresponding Author

Abstract

The institutional capacity to deal with crime data presents a significant challenge for law enforcement agencies. This study develops and evaluates a machine learning (ML)-based crime detection system based on four supervised classification algorithms—Decision Tree, Logistic Regression, Random Forest, and K-Nearest Neighbour (KNN)—applied to a publicly available crime dataset of 70,894 records obtained from Kaggle. Important spatiotemporal features like year, day of week, hour, latitude, and longitude were used as independent factors to anticipate the offence code as the dependent variable. R2 score, accuracy, and Mean Absolute Error (MAE) were employed.The results revealed that the Random Forest Classifier had the highest accuracy of 94% and the lowest MAE of 1,024, outperforming the Decision Tree (86%), Logistic Regression (71%), and KNN (57%). These findings lend credence to the notion that tasks requiring the classification of crime categories are best suited for ensemble-based techniques. The proposed system has a good potential of being used by Nigerian law enforcement agencies as a progressive, data-driven tool to support real-time crime detection and prevention.

Keywords: *Crime Detection, Machine Learning, Random Forest, Offence Classification, Predictive Policing*

1. INTRODUCTION

The safety, security, and economic growth of societies spanning from tiny rural settlements to huge metropolitan centers are all at risk due to crime, which continues to be one of the most enduring and complicated issues facing civilizations worldwide. Law enforcement organizations tasked with upholding public safety and order are under increasing pressure as a result of the diversification and intensification of criminal activity brought on by the acceleration of urbanization and the deepening of socioeconomic disparities. More efficient, data-driven methods to crime management are desperately needed in Nigeria and many other developing countries because the responsibility of crime detection and avoidance is increasing disproportionately to the institutional capability of security services [1].

In the past, investigation methods, eyewitness accounts, and the experience of law enforcement officers were all crucial to the detection of crimes. Although these techniques are still fundamental, it is becoming more and more clear how inadequate they are for managing massive amounts of crime data and forecasting criminal trends. New avenues for computational techniques to crime analysis have been made possible by the quick spread of digital records, standardized surveillance tools, and open crime datasets. Machine learning (ML) has become known as a game-changing technology in artificial intelligence (AI), facilitating the automatic discovery of patterns and correlations in complex datasets that human analysts might otherwise overlook.[2].

Recent advances in machine learning have demonstrated tremendous potential for crime detection and prediction. Supervised learning methods like K-Nearest Neighbor (KNN), Random Forest, Logistic Regression, and Decision Tree classifiers have been widely used on crime datasets with various degrees of success. Specifically, ensemble-based methods have shown improved predictive performance by merging the output of multiple classifiers to produce more accurate and dependable predictions. Studies conducted in a variety of institutional and geographic contexts have shown that Random Forest regularly outperforms single-classifier approaches for handling high-dimensional, imbalanced crime datasets. [3], [4]. Despite these developments, comparably few research have applied these types of algorithms in Nigeria, where unique opportunities and limitations are presented by institutional structures, crime patterns, and data quality.

An increasing body of research emphasizes the need of model interpretability and explainability for crime prediction systems. Explainable artificial intelligence (XAI) frameworks are increasingly being promoted to ensure that law enforcement personnel and policymakers can understand, audit, and trust the predictions made by machine learning algorithms [5]. If they are not interpretable, even highly accurate models have the potential to undermine public confidence and propagate systemic biases found in historical crime statistics. This is especially important in developing countries when algorithmic accountability systems are still in their infancy. Therefore, the aim of this research is to develop and evaluate a crime detection system by using four supervised machine learning algorithms—Decision Tree, Logistic Regression, Random Forest, and K-Nearest Neighbors—on a crime dataset that is available to the public. By carefully evaluating the performance of several algorithms utilizing accuracy, Mean Absolute Error (MAE), and R^2 score, this work seeks to identify the optimal model for crime prediction and provide helpful information for law enforcement authorities in Nigeria and similar environments. The remainder of the paper is structured as follows: Section 2 reviews the pertinent literature; Section 3 describes the materials and methodology; Section 4 presents the model implementation and results; and Section 5 offers recommendations for more research.

2. RELATED WORKS

A deployment-ready machine learning model for predictive analysis of rising crime patterns is presented by Halder et al. [3]. To improve class balance and model accuracy, the study integrated data augmentation techniques such as SMOTE and feature scaling with Gradient Boosting, Decision Tree, Random Forest, and Support Vector Machine (SVM) classifiers. Notably, SVM achieved the highest accuracy of 0.96 when SMOTE and feature scaling were coupled, highlighting the critical role preprocessing plays in improving ML crime prediction abilities. The authors concluded that the careful use of augmentation strategies can significantly increase the potential for generalization of crime prediction models.

In an AI-driven crime prediction framework, Madubuike et al. [2] evaluate the accuracy, precision, recall, F1-score, ROC-AUC, and calibration metrics of Logistic Regression, SVM, Random Forest, Gradient Boosting, and Deep Neural Networks using historical data analysis. The findings confirmed that ensemble methods and gradient boosting yielded superior outcomes. It provided a scalable and repeatable method for sophisticated crime analytics systems which may be utilized by law enforcement agencies in Nigeria and other poor countries, in addition to resolving fairness and privacy concerns.

LCRNet, a lightweight deep learning framework for identifying crime patterns, is introduced by Lu et al. [6]. It integrates a Convolutional Neural Network (CNN) with a transformer architecture and simulated annealing sparsity. When tested on real criminal activity in Los Angeles, the model had an accuracy of 97.76% and significant cross-dataset generalizability. The study represents a major development in the application of deep learning to crime investigation and demonstrates how hybrid transformer-CNN architectures have the potential to beat conventional ML techniques in terms of accuracy and efficiency.

A comprehensive review of the literature on AI-driven crime prediction was conducted by Iqbal et al. [1]. By combining results from a variety of sources, the study demonstrated the advantages of ensemble learning methods, particularly Random Forest, in achieving high projected accuracy. Transparency, interpretability, and ethical fairness must be addressed by future systems, the authors stressed, particularly when they are used to guide important law enforcement choices in resource-constrained settings. Additionally, they identified a crucial weakness in the explicit integration of Explainable AI (XAI) methods into crime prediction models.

Using bagging and boosting ensemble algorithms, Sani et al. [4] developed an enhanced crime prediction model that was used on both an international benchmark dataset and the Nigerian crime dataset in Kano State. The study examined classifiers using Decision Tree, Random Forest, Gradient Boosting, and Extreme Gradient Boosting. On the Nigerian dataset, all ensemble models achieved accuracy scores of 88%, suggesting that boosting-based techniques offer substantial benefits over standalone classifiers. The findings demonstrate how ensemble techniques can be adapted to African crime data scenarios and point to meta-learning and dynamic

ensemble selection as possible directions for future performance improvement.

The effectiveness of crime investigation is analyzed by Shivaprasad et al. [7]. They adopted a strategy that employs machine learning to detect crimes. The predictions were made using the K-Neighbor classification technique and multilinear regression. Their model was evaluated and found to have an accuracy score of 85%, which was 25% higher than previous researches

A survey on crime incidence, detection, and prediction methods is conducted by Shruti and Shraddha [8]. Their survey's objective was to investigate data mining methods that could aid in crime detection. The American Fact Finder 2014 dataset was utilized. Numerous methods, including K-means clustering, decision trees, Naive Bayes, apiriori algorithm, ANN, and DNN, were employed. DNN was shown to have the best accuracy score for multimodal data fusion, at roughly 84.25%.

In an effort to predict and prevent crime, Shah et al. [9] investigate crime forecasting with an emphasis on machine learning and computer vision. Their study's main goal was to investigate how the crime rate negatively affects social interactions. Both quantitative and qualitative methods were applied. For their research, algorithms including K-Nearest Neighbors, Pearson-R correlation, and linear regression were employed. K-Nearest Neighbor achieved the highest overall precision score among the three algorithms.

Kshatri et al. [10] use an ensemble approach and stacked generalization to empirically analyze machine learning algorithms for crime prediction. The study's datasets were sourced from Vancouver's open data catalog. enhanced decision tree and K-nearest-neighbor. In Vancouver, crime prediction accuracy ranged from 39% to 44%.

Sritha et al. [11] use data mining to create an overview of a crime detection system. The model to forecast crime in a particular location and at a particular time is the main topic of the study. Crime pattern detection was the methodology employed, and SQL injections are used in the system. It is a query-based technique for identifying SQL attacks and other online threats. After examining all of these factors, the study arrived at its primary objective, which might have an impact on crime detection in any nation.

3. MATERIALS AND METHODS

This section presents the research methodology used in the paper. All procedures that were taken to achieve the desired results are shown in this section. These include data collection, development of ML models, training and testing the datasets, model's prediction and evaluation. Figure 1 depicts the proposed crime detection system architecture of the of the study.

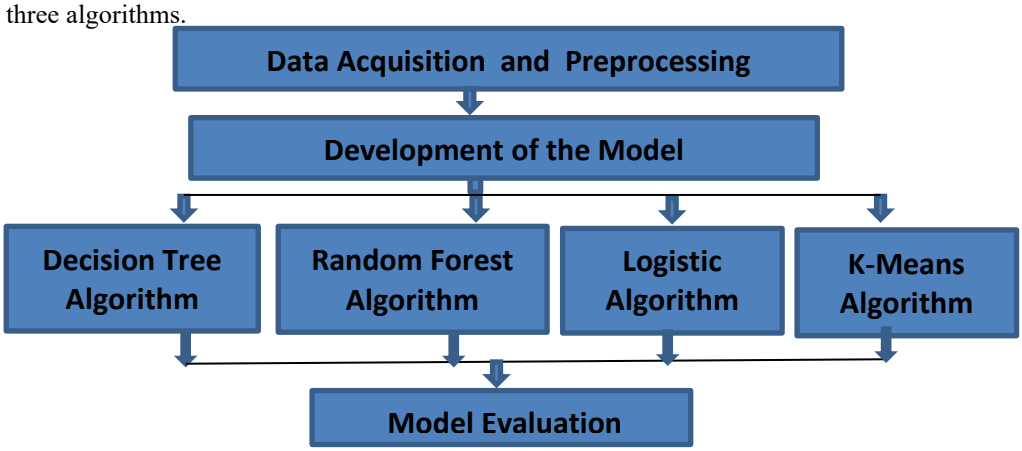


Figure 1: Architecture of the Crime Detection Model

3.1 Data Acquisition and Preprocessing

The dataset used for the implementation of crime detection model was downloaded from the online data platform called Kaggle. This dataset was saved in CSV file formats.

The dataset was divided into two parts: Testing and Training data at the ratio of 80% and 20% respectively. The dataset has 70,894 rows and 17 columns. Dataset was imported using the Pandas library function. The key features such as Year, Day of Week, Hour, Incident Number, Lat and Long are selected from the dataset as the model input features or independent variables. The feature or label which is the Offense Code is selected as the model's output also known as dependent variable (Figure 2).

1 Exploratory data analysis

In [4]: `df.describe()`

Out[4]:

	INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	SHOOTING	YEAR	MONTH	HOUR	UCR_PART	Lat	
count	7.089400e+04	70894.000000	0.0	70894.000000	70894.0	70894.000000	70894.000000	0.0	70894.000000	7.089400
mean	2.020772e+08	2292.077200	NaN	0.015826	2020.0	6.595878	12.815979	NaN	41.278263	-6.933145
std	3.830416e+06	1257.571021	NaN	0.124805	0.0	3.452960	6.438212	NaN	6.563217	1.102357
min	1.906588e+07	111.000000	NaN	0.000000	2020.0	1.000000	0.000000	NaN	0.000000	-7.117867
25%	2.020247e+08	801.000000	NaN	0.000000	2020.0	3.000000	9.000000	NaN	42.295129	-7.109835
50%	2.020493e+08	3005.000000	NaN	0.000000	2020.0	7.000000	14.000000	NaN	42.325303	-7.107671
75%	2.020731e+08	3125.000000	NaN	0.000000	2020.0	10.000000	18.000000	NaN	42.346735	-7.105978
max	8.546523e+08	99999.000000	NaN	1.000000	2020.0	12.000000	23.000000	NaN	42.395042	5.250001

Figure 2: Dataset used for the Model

3.2 Algorithms for the Proposed Model

Among the several classification algorithms currently in use, four machine learning algorithms are employed: decision trees, logistic regression, random forests, and K-neighbors. All these algorithms in the model were and implemented on Python's Sklearn libraries.

In order to implement the model accurately, the data had to be preprocessed until irrelevant, incomplete, inconsistent and null values were removed from the raw dataset with 70,894 records. First, the missing values of records in the dataset were checked where the incomplete records were dropped. Second, no outlier or null values present in each record to avoid a possible decline in the accuracy of the model prediction.

a. Decision Trees Classifier

Decision Trees display the predictions that come from a sequence of feature-based splits using a flowchart that resembles a tree structure. It starts from a root node and ends with a choice made by leaves. The root node, decision node, and leaf node are its three constituent parts. A root node is the node at which population division begins. Decision nodes are those that are obtained by splitting a root node, and leaf nodes are those that cannot be further split. Equation 1's formula is used to fit the Decision Tree Classifier's mathematical model.

$$Gini\ Index = 1 - \sum_{i=1}^{n} (Pi)^2 \qquad (1)$$

"Success" or "failure" is the categorical target variable that the Gini index uses. It only does binary splits. Greater inequality and heterogeneity are implied by a higher Gini index score. Assigning success (p) and failure (q) ($p^2 + q^2$) and calculating the Gini index for split using the weighted Gini score of each node of that split are the steps involved in utilizing the Gini index to train the machine learning model.

b. Logistic Regression

The linear regression equation can be used to create the logistic regression model. Equation 2 below provides the mathematical procedures to obtain the logistic regression equations.

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n \qquad (2)$$

The range of y in logistic regression is limited to 0 and 1. Equation (2) is therefore divided by (1-y), as Equation 3 illustrates.

$$\frac{y}{1-y} = 0 \qquad (3)$$

for y = 0, and infinity for y = 1 (3)

However, the required range is –[infinity] to +[infinity]. After that, the equation's logarithm will become:

$$\text{Log}\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n \qquad (4)$$

This is the final equation for Logistic Regression. The mathematical model of the Logistic Regression is fitted using the formula as shown below:

$$\text{Log}\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n \qquad (5)$$

The 'y' represents the dependent variable, while the 'x' represents the independent variables or features, which can be as many as needed. For this ML model, six (6) independent variables ('day of the week', 'hour', 'incident number', 'lat', 'long' and 'year') are used.

$$\text{Log}\left[\frac{y}{1-y}\right] = b_0 + b_1 \text{ day of week} + b_2 \text{ hour} + b_3 \text{ incident number} + b_4 \text{lat} + b_5 \text{long} + b_6 \text{year} \quad (6)$$

where 'b' represents the non-standardized regression weight. It is a measure to multiple linear regression weight can be simplified in its interpretation. For example, as variable 'hour' increases, the likelihood of getting a perfect prediction for y should also increase.

c. Random Forest Classifier

The Random Forest algorithm is an effective tree learning technique in machine learning. It produces a large number of Decision Trees during the training phase. Each tree is constructed using a randomized subset of the data set in order to identify a random selection of attributes in each division. The randomization makes each tree more varied, which reduces the possibility of overfitting and improves overall prediction performance. For regression tasks, the method averages the output of each tree; for classification tasks, it votes. The findings of this collaborative decision-making process, which benefits from multiple trees' insights, are reliable and consistent. Random forests are widely used for regression and classification problems because they can manage complex data, reduce overfitting, and generate reliable forecasts in a range of circumstances.

The formula for calculating the feature importance is given in Equation 7.

$$Fi = \frac{\sum_j node\ j\ splits\ on\ feature\ ni_j}{\sum_{k \in all} node\ ni_k} \quad (7)$$

d. KNN

Using the dataset, we have the first value of month and hour features as 2 and 0 respectively. Then, values assigned to the variables x_1 and x_2. Suppose the value of k = 3, calculating the distance between the query instance and all the training samples using the formula in Equation 8.

$$(x_2 - x_1)^2 + (x_2 - x_2)^2 = Y \quad (8)$$

Doing this for all the features in the datasets, the distance and determine nearest based on the k-th minimum distance is sorted. The prediction would then be the simple majority of values; the highest number of occurrences of a particular value would be the predicted outcome of the model.

3.3 Performance Evaluation Metrics

The performance of the model developed will be assessed using the standard metrics: Mean Absolute Error, R^2 Error, and Accuracy score for each of the algorithm.

a. Mean Absolute Error (MAE)

Without taking into account the orientations of the errors, this metric calculates the average size of the errors in a set of forecasts. It evenly calculates the average of each unique difference. Equation 9 shows the formula of this measure.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \frac{(O_i - F_i)}{O_i}\ x\ 100 \quad (9)$$

where O_i refers to the actual crime committed, and F_i refers to the predicted crime and "n" refers to the total number of crimes committed.

b. Accuracy Score

One parameter used to assess classification models is accuracy. Roughly speaking, accuracy is the percentage of correct predictions our model made. The following is an illustration of its formula in Equation 10:

$$Accuracy\ Score = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (10)$$

c. R^2 Score

The R^2 score is a very important metric that is used to evaluate the performance of a regression-based ML model, otherwise known as the coefficient of determination. It works by measuring the amount of variance in the predictions explained by the dataset. The correlation coefficient can be calculated as shown in Equation 11.

$$r = \frac{n(\sum xy) - \sum x \sum y}{\sqrt{[n^*(\sum x^2 - (\sum x)^2)] * [n^*(\sum y^2 - (\sum y)^2)]}} \quad (11)$$

where r = The Correlation coefficient, n = number in the given dataset, x = first variable in the context, and y = second variable.

Then:

$$R^2 = 1 - \frac{sum\ square\ regression}{total\ sum\ of\ square} = 1 - \frac{\sum(y_i - y)^2}{\sum(x_i - x)^2} \quad (12)$$

where the sum squared regression is the sum of the residuals squared, and the *total sum of squares* is the sum of the distance the data is away from the mean all squared. As it is a percentage, it will take values between 0 and 1 as shown in Equation 12.

4. MODEL IMPLEMENTATION

The model uses the dataset of crimes that have been committed in the past, in return it would be able to learn from it and predict future crimes that are likely to happen based on the data it learned from. For this model, the study used dataset of crimes from the year 2020. The dataset is shown in Figure 3.

In [53]: df

Out[53]:

	INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE
0	854652314	3115	NaN	INVESTIGATE PERSON	B2	289	0	20-02-20 0:00
1	457856954	3115	NaN	INVESTIGATE PERSON	B2	289	0	21-02-20 1:42
2	302030654	1106	NaN	FRAUD - CREDIT CARD / ATM FRAUD	D4	167	0	24-04-20 13:00
3	212025777	1108	NaN	FRAUD - WELFARE	A7	28	0	1/5/2020 0:00
4	212025557	2670	NaN	HARASSMENT/ CRIMINAL HARASSMENT	A1	77	0	7/9/2020 0:00
...	...	...	...	...	...	...	...	...
19	20063425	3114	NaN	INVESTIGATE PROPERTY	A7	21	0	1/9/2020 0:00
10	20062358	1107	NaN	FRAUD - IMPERSONATION	E18	520	0	28-08-20 18:39
11	20046400	1501	NaN	WEAPON VIOLATION - CARRY/ POSSESSING/ SALE/ TR...	B2	330	0	2/7/2020 1:38
12	20038446	1501	NaN	WEAPON VIOLATION - CARRY/ POSSESSING/ SALE/ TR...	B2	300	0	3/8/2020 1:15
13	20030892	540	NaN	BURGLARY - COMMERICAL	C11	380	0	3/5/2020 0:00

14 rows × 17 columns

Figure 3: Imported the Dataset using Pandas Library

4.1 Data Exploration and Preprocessing

Data can either be structured and unstructured. For this model, the study performed exploratory data analysis and preprocessing which looked at the data in a more detailed view in order to know if there are null values and outliers (Figure 4). The figure shows that the dataset has some null values. For example, the total number of rows is 70,894. Any column that is showing number less than 70,894 means there are null values (empty data cells) in it. For this model, no reason to remove the null values, because the features used for the model have no null values.

```
In [5]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70894 entries, 0 to 70893
Data columns (total 17 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   INCIDENT_NUMBER       70894 non-null  int64
 1   OFFENSE_CODE          70894 non-null  int64
 2   OFFENSE_CODE_GROUP    0 non-null      float64
 3   OFFENSE_DESCRIPTION   70894 non-null  object
 4   DISTRICT              70616 non-null  object
 5   REPORTING_AREA        70894 non-null  object
 6   SHOOTING              70894 non-null  int64
 7   OCCURRED_ON_DATE      70894 non-null  object
 8   YEAR                  70894 non-null  int64
 9   MONTH                 70894 non-null  int64
 10  DAY_OF_WEEK           70894 non-null  object
 11  HOUR                  70894 non-null  int64
 12  UCR_PART              0 non-null      float64
 13  STREET                70893 non-null  object
 14  Lat                   70894 non-null  float64
 15  Long                  70894 non-null  float64
 16  Location              70894 non-null  object
dtypes: float64(4), int64(6), object(7)
memory usage: 9.2+ MB
```

Figure 4: Data Preprocessing

4.2 Data Feature Selection

Data feature selection is all about extracting all the necessary features (columns) that contain independent variables and the dependent variable. In ML, the model works using the formula in Equation 13.

$$X + Y = Z \qquad (13)$$

X and Y are the independent variables or features that would determine the outcome of the model which is Z (dependent variable). Data feature selection entails selecting from the data sets all the features (columns) that would help the model to predict a crime. So, the features were extracted as shown in Figure 5 while a new view is shown in Figure 6.

3) Data Feature Selection

```
In [9]: # Extracting Features

dataset = pd.DataFrame()
dataset['DAY_OF_WEEK'] = df['DAY_OF_WEEK']
dataset['HOUR'] = df['HOUR']
dataset['INCIDENT_NUMBER'] = df['INCIDENT_NUMBER']
dataset['Lat'] = df['Lat']
dataset['Long'] = df['Long']
dataset['YEAR'] = df['YEAR']
dataset['OFFENSE_CODE'] = df['OFFENSE_CODE']
```

Figure 5: Data Feature

```
In [10]: dataset.head()
Out[10]:
```

	DAY_OF_WEEK	HOUR	INCIDENT_NUMBER	Lat	Long	YEAR	OFFENSE_CODE
0	Thursday	0	854652314	42.333679	-71.091878	2020	3115
1	Friday	1	457856954	42.333679	-71.091878	2020	3115
2	Friday	13	302030654	42.339542	-71.089409	2020	1106
3	Friday	0	212025777	42.370818	-71.039291	2020	1108
4	Monday	0	212025557	42.361839	-71.059765	2020	2670

Figure 6: New View of Data

From this dataset, dependent variable Z was extracted which served as the output or the predicted value of our model as the OFFENCE_CODE. The OFFENCE_CODE has unique integer values that tell of the specific offence committed by a person. From the dataset, the crime of "Credit Card/ATM Fraud" has a unique offence code of 1106. Anytime the model predicts the numeric value 1106, it implies that it is predicting a crime of "ATM/Credit Card Fraud".

4.3 Categorical Values

Categorical values are values that a ML model cannot process, which are called strings. For this model to work, there is a need to convert some string values to integer using the "fit.transform" function. The column "DAYS_OF_THE_WEEK" has string values.

So, using fit.transform, it assigns a unique number to each day of the week in order for the ML model to process it (Figure 7).

```
In [12]: # Dealing with categorical features
         from sklearn.preprocessing import LabelEncoder

In [13]: le = LabelEncoder()

In [14]: dataset.iloc[:,0] = le.fit_transform(dataset.iloc[:,0])

In [15]: # We have to separate the features and labels
         x = dataset.iloc[:,:-1]

In [16]: dataset.sample(10)
Out[16]:
```

	DAY_OF_WEEK	HOUR	INCIDENT_NUMBER	Lat	Long	YEAR	OFFENSE_CODE
11692	1	4	202080922	42.341824	-71.055503	2020	3801
8268	4	6	202085337	42.282951	-71.093775	2020	3410
33287	4	21	202052258	42.354419	-71.150188	2020	614
37911	5	21	202046057	42.355484	-71.137418	2020	3005
10941	4	15	202081850	42.259365	-71.159330	2020	3802
60306	0	19	202014788	0.000000	0.000000	2020	3831
61832	0	12	202012710	42.333805	-71.093778	2020	3810
5235	0	22	202089326	42.347757	-71.138563	2020	3006
53325	4	10	202024540	42.315043	-71.111898	2020	801
14996	2	20	202076656	42.313224	-71.057153	2020	801

Figure 7: Categorical Value of Dataset

a. Separating the Independent Variables from the Dependent Variable

At this stage, we separated our X and Y from Z which are the independent variables (features) from the dependent variable i.e. predicted outcome (Figure 8).

```
In [17]: # Separate the features and labels

         x = dataset.iloc[:,:-1]
         y = dataset.iloc[:,-1]

In [18]: x.head()
Out[18]:
```

	DAY_OF_WEEK	HOUR	INCIDENT_NUMBER	Lat	Long	YEAR
0	4	0	854652314	42.333679	-71.091878	2020
1	0	1	457856954	42.333679	-71.091878	2020
2	0	13	302030654	42.339542	-71.089409	2020
3	0	0	212025777	42.370818	-71.039291	2020
4	1	0	212025557	42.361839	-71.059765	2020

```
In [19]: y.head()
Out[19]: 0    3115
         1    3115
         2    1106
         3    1108
         4    2670
         Name: OFFENSE_CODE, dtype: int64
```

Figure 8: Variables Separation

Algorithms	Metrics		
	Accuracy (%)	MAE	R^2 score
Decision Tree Classifier	86%	1189	-0.77
Logistic Regression	71%	1053	-0.48
Random Forest Classifier	94%	1024	-0.5
KNN	57%	1405	-1.23

4.4 Results of Model Implementation using the Algorithms

The Table 1 shows the results of the implementation of the four ML algorithms on the Sklearn library of Python programming language.

Table 1: Model Performance Evaluation

b. Splitting the Dataset into Testing and Training datasets

For the ML model to be accurate, some percentage of the dataset needs to be allocated for training, while the rest for testing. For this model, we assigned 80% (Percent) of the datasets for training, while the remaining 20% for testing (Figure 9).

```
In [20]: # Splting the data into training and testing dataset

         from sklearn.model_selection import train_test_split

In [21]: x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.4,random_state=2)
```

Figure 9: Datasets Splitting Running Layer

From all the algorithms we have used so far to test our model, the algorithm with the highest accuracy score is Random Forest Classifier (95.8%), and the lowest is KNN algorithm (57%).

The performance evaluation best suited for this model in terms of accuracy result and Absolute Mean Error Score is Random Forest Classifier as shown in Table 1. Figures 10-13 show the graphical outputs of the performance evaluation of the algorithms.

4.3. Model Performance Evaluation

```
In [55]: df
Out[55]:
```

	ALGORITHM	ACCURACY SCORE
0	LOGISTIC REGRESSION	71
1	RANDOM FOREST	94
2	DECISION TREE	86
3	KNN ALGORITHM	55

Figure 10: Model Evaluation

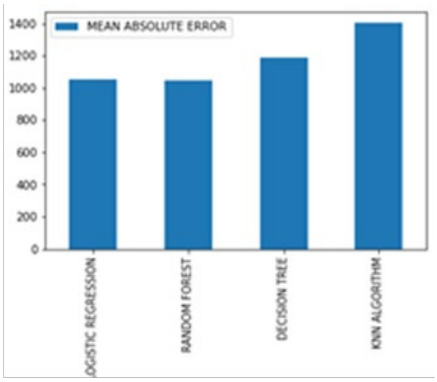


Figure 11: Graphical MAE of the Algorithms

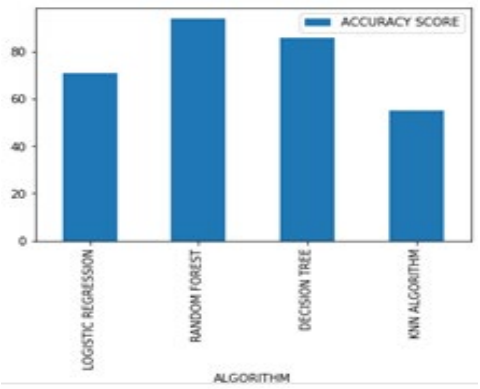


Figure 12: Graphical Accuracy Score of the Algorithms

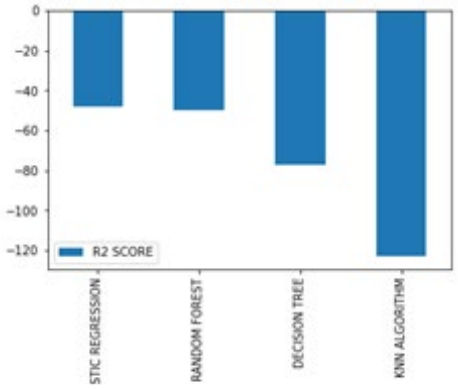


Figure 13: Graphical R^2 Score of the Algorithm

This implies that the algorithm best suited for this model in terms of accuracy score and Absolute Mean Error Score is Random Forest Classifier.

4.4 Discussion of Findings

Using a combination of spatiotemporal and calendar-based variables, this study provides a crime detection system that uses machine learning algorithms to analyze past crime data and predict the kind of offenses that are likely to occur. The Decision Tree Classifier, Logistic Regression, Random Forest Classifier, and K-Nearest Neighbor (KNN) algorithms were evaluated.

Across all evaluations, the Random Forest model performed the best, supporting a growing body of recent research showing the effectiveness and reliability of ensemble learning techniques in predictive crime modeling.

The Random Forest Classifier had the best predictive performance out of all the models examined, with an accuracy of 94% and the lowest Mean Absolute Error (MAE) of 1024. This is explained by its ensemble architecture, which reduces the danger of overfitting that often plagues individual tree-based models by combining the outputs of multiple decision trees. The results are in close agreement with those of [4], who found ensemble models reaching 88% accuracy on a Nigerian crime dataset, and with [3], in which Random Forest ranked among the leading classifiers prior to data augmentation pushing SVM performance to 0.96. The Decision Tree Classifier proved competitive as well, attaining 86% accuracy, whereas Logistic Regression reached 71%. KNN recorded the weakest result at 57%, which can be explained by its known vulnerability to high-dimensional input spaces and its dependence on proximity-based measures that may fail to capture the intricate nonlinear patterns present in crime data.

Negative R2 ratings were found for all four algorithms, meaning that none of the models performed better than a straightforward mean-based baseline under regression evaluation criteria. However, considering the categorical nature of the target variable (OFFENCE_CODE) and the dataset's significant class imbalance, this is an expected result. Therefore, rather than using regression-oriented metrics like R2, this discovery emphasizes the necessity to assess classification models using accuracy and confusion-matrix-based metrics. Future research could address class imbalance and improve each model's MAE and predictive consistency by using oversampling techniques like SMOTE, as employed in [3].

Transparency and responsible use of predictive models are significant issues brought up by the larger body of literature. Explainable AI (XAI) technologies are not often used in crime prediction frameworks, according to Iqbal et al. [1] and a concurrent systematic review from both. Despite achieving excellent classification accuracy, the current model operates as a "black box," producing predictions without providing comprehensible insight into the characteristics that influence each result. Law enforcement personnel are unlikely to rely on forecasts they cannot comprehend or examine in operational settings. Therefore, it is recommended that future iterations of this model include XAI techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), which can offer localized explanations of feature contributions and thereby improve the system's ethical accountability and practical utility. This factor is especially important when the model is used to direct proactive policing, an area where accountability, fairness, and transparency are unavoidable.

The study's input characteristics, which include year, day of the week, hour, incident number, latitude, and longitude, are essentially spatiotemporal and place the model in a predictive police framework that is focused on locating crime hotspots. Advanced deep learning architectures, such as those based on transformer models and convolutional neural networks, can provide accuracy rates above 97% and extract deeper spatial and temporal representations from crime data, according to research by Lu et al. [6]. These models represent a reasonable path for future improvement, even though they require far more processing power and larger training datasets than those utilized in this study. For comparison with more sophisticated methods, the current work thus offers a reliable and repeatable baseline. A systematic comparison of Random Forest with deep learning architectures and gradient boosting techniques utilizing Nigerian crime datasets would be a top priority for future study.

At the social policy level, the implementation of machine learning methods for crime detection in Nigeria must be explored in tandem with careful consideration of data governance, institutional preparedness, and the need to foster public trust. According to Madubuike et al. [2], explainability should be treated as a basic design need that is integrated from the beginning rather than as an optional addition in crime prediction systems meant for real-world implementation. The system created in this study represents a significant initial step toward an intelligent crime detection approach that can assist Nigerian law enforcement in preventative, proactive security operations. However, in order to fully realize its potential, it will be necessary to connect to real-time crime data streams, continuously retrain the model as new data becomes available, and engage with policymakers and community stakeholders to ensure that deployment remains both efficient and ethically acceptable.

```
In [71]: df
Out[71]:
```

	ALGORITHM	ACCURACY SCORE
0	LOGISTIC REGRESSION	72.9
1	RANDOM FOREST	95.8
2	DECISION TREE	86.1
3	NAÏVE BAYES	60.0
4	KMEANS ALGORITHM	54.6
5	KNN ALGORITHM	57.0

Figure 14: Graphical Comparison of Algorithms

5.0 CONCLUSION

The study of crime detection system using ML model as done in this paper, has become imperative in any nation's security especially in the developing nation like Nigeria due to its attendant benefits to mitigate security challenges toward ensuring a safe society. Many studies adopted different approaches to the area of crime prediction and also compared one with each other. The findings of this study showed random forest algorithm performed better among other algorithms used. It can be concluded that the comparison of different algorithms for crime prediction system in model building give some indications, such as the selection of an algorithm may depend on the dataset type (like locations where the crime happens, type of crime committed and year, etc.), and random forest algorithm performed well on the dataset. It is therefore recommended for law enforcement agencies in Nigeria to make use the proposed model for the purpose of real-time detection of these crimes before they are even committed. This will afford security operatives to keep residents abreast of security-related information in their geographical location. Also, the outcome of the research work is recommended for researchers in this field of study for future guide while future research can be conducted on other sources of datasets.

REFERENCES

[1] Iqbal, N., Hassan, A., & Waheed, T. (2025). AI-driven crime prediction: a systematic literature review. Journal of Computational Social Science, 8(2), 1–47. https://doi.org/10.1007/s42001-025-00373-z

[2] Madubuike, C. L., Nosike, P., & Ezeonyi, N. (2026). Artificial Intelligence-Driven Crime Prediction Through Historical Data Analysis and Machine Learning Models. Global Journal of Artificial Intelligence and Technology Development. https://doi.org/10.xxxxx

[3] Halder, M., Sinha, P., Biswas, D., Yang, T., & Rathore, R. S. (2026). Deployment of Machine Learning Model for Predictive Analysis of Rising Crime Trends. In A. Swaroop, V. Kansal, & A. E. Hassanien (Eds.), Proceedings of Sixth Doctoral Symposium on Computational Intelligence (DoSCI 2025). Lecture Notes in Networks and Systems, Vol. 1498. Springer, Singapore. https://doi.org/10.1007/978-981-96-8104-4_27

[4] Sani, J. A., et al. (2025). Enhanced Crime Prediction Model Using Bagging and Boosting Algorithms. DUJOPAS, 11(4b), 75–82.

[5] Jamadar R., Deshmukh K., Deshmukh, N., Kadale A. and Padwal A. (2024). Comparative Study of Ml-Based Crime Prediction System. International Journal of Research Publication and Reviews, 5(2), 3451-3459.

[6] Lu, H., Chen, C., & Ma, Y. et al. (2025). Lightweight deep learning model for crime pattern recognition based on transformer with simulated annealing sparsity and CNN. Scientific Reports, 15, 20311. https://doi.org/10.1038/s41598-025-07260-7

[7] Shivaprasad, M., Sakshi, M., Saloni, K., & Hafsa, B. (2021). Crime Prediction Using Machine Learning Approach. *International Journal of Advanced Research in Computer and Communication Engineering,* 10(5), 23-31.

[8] Shruti, S. & Shraddha, S. (2018). A Survey on Crime Occurrence Detection and Prediction Techniques. *International Journal of Management, Technology and Engineering,* 8(12): 45-56.

[9] Shah N., Bhagat N. and Shah M. (2021). Crime Forecasting: A Machine Learning and Computer vision Approach to crime prediction and prevention. Vis. Comput. Ind., Biomed., Art; 4(1): 1–14.

[10] Kshatri S. S., Singh D., Narain B., Bhatia S., Quasim M. T., and Sinha G. R. (2021). An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: An ensemble approach. IEEE Access, 9,: 67488–67500.

[11] Sritha Z., Digvijay, P. & Ismail, S. (2020). An overview of a crime detection system using the art of data mining. *International journal of innovations in engineering research and technology*, 7(5); page 125-128.